the incidence of liver & lung and Bronchus & tongue & lips cancer with number of smoker in USA

A Data Management Plan created using DMPonline

Creator: khaled maher

Affiliation: Other

Funder: European Commission (Horizon 2020)

Template: DMP University of Vienna English V2

ORCID iD: 0000-0002-6316-9087

Project abstract:

this dataset is used to show the relation between number of smoker in united state and the number of incidence of (liver , lunge &bronchus , lips , tongue) cancer

Last modified: 18-04-2019

the incidence of liver & lung and Bronchus & tongue & lips cancer with number of smoker in USA

1. Administrative Data

the relationsahip between the liver cancer incidance and tobacco smoker in alabama state 2000-2010

name : Khaled Maher Yahia Awad
tel : 00972592496300
e-mail : khaledmaher024@gmail.com

tel : 00972592496300
e-mail : khaledmaher024@gmail.com

version : 1 (first version)
date : 16/4/2019

2. Data Collection

- csv file that contain attribute like (state : string , year : int64 , count of liverCancerIncidance : int64 , count of TobaccoSmoker : int64)
- 371 bytes

the data had been collected from 2 datasets the first one was the liver cancer incidance that collected from data.gov website and the second one (tobacco usage) was collected from kaggle website
the data had been produced using a set of Python functions that make some transformation on the 2 original datasets

3. Documentation

the final documentation is collected and created to describe the effect of increasing number of tobacco smoker in one of USA states on the increasing number of people who is suffering from liver cancer , the data collected by the author : khaled maher awad , and its collected from 2 datasets from different websites , the first one is from kaggle and its about number of tobacco usage in USA state from 1995 - 2015 and the other one was collected from data.gov website and its about liver cancer incidance in USA states from 1999 to 2010 , i made a set of procedure on the original datasets to prepare and analyze the data to be useful and can be used , first of all is to check if the data has null values or not and fill the null field with the median value of the column , the second procedure that eleminate the rows that is out of the specified period (2000-2010) , then i convert the type of one column in the tobacco usage "Never smoked" by deleting the percentage mark and change the results string into float , then i calculated the number of smoker by substracting 1 by the value of "Never smoker" column , and also i eliminated the rows that doesnt represent liver cancer in the second dataset , and i make a function that work as groupby function and i added a new column that represent the summation or total # of incidance in one year for all years in the specified period , then i merged the 2 datasets to results the final result

4. Metadata

the metadata structured by xml file that contain the basic information about the datasets like (title , creator , subject , description , date , type , format , source , language , coverage , rights)

5. Ethics and Legal Compliance

in the research and collecting data we put in account that there isnt any ethically questionable material included just in Tobacco Usage dataset there is a coulmn about sex but i think it will not be a big deal or it will not harm any person , there is no limitation on the image size or resolution just that to be enough to include all person in USAno other country , the dataset is open access and with no limitation

The resulted dataset is open access to all users to access , maintain , use , share and so on , the original datasets thave license of ODBL but the other one has a license but its unknown

6. Storage and Backup

the final result is stored as a CSV file in the computers hard drive and the data will be secured and backed up using barracuda back up service (https://www.barracuda.com) and the data will be backed up weekly to prevent any data loss

the final data has no sensitive data and all the data has an open access "all people can access the final result" ,there is no risk on the data being attacked because there is no personal data

7. Selection and Preservation

the data that is wanted to be shared for long time is the final / resulted data that include the state name , the year , number of smoker , number of liver cancer incidance and these data is available as csv file , the data has no specified time to be stored in the repository
the presistent identifier for the data is :
10.1234/khaled.maher

the final data is shared at github website in the repository SW_Exercise and the address : https://github.com/KhaledMaher024/SW_Exercise and published at Zenedo website and has the name of : correlation between liver cancer incidance and tobacco smoker in alabama 2000-2010
with address : https://zenodo.org/record/2642043
there is no cost on carging additional data

8. Data Sharing

the data will be found on github and zenedo websites , there is no restriction on the access on the data , the license of the reulted data is (Creative Commons Attribution 4.0 International)

the data visualized as a flowchart and it can be used as an excel file also

9. Responsibilities and Resources

the responsible for implemeinting the DMP is khaled awad , and also the responsible on ensuring that the DMO is reviewed and revised

no